Relevance Feedback in Deep Convolutional Neural Networks for Content Based Image Retrieval

Maria Tzelepi
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
mtzelepi@csd.auth.gr

Anastasios Tefas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
tefas@aiia.csd.auth.gr

ABSTRACT

In this paper a novel Relevance Feedback approach that uses deep Convolutional Neural Networks (CNNs) for image retrieval is proposed. We utilize a deep CNN model to refine the feature representations of the deeper layer used for the retrieval, based on the feedback of the user. To this end, we adapt the pretrained model and we re-train the corresponding neural layers on relevant and irrelevant images, as qualified by the user. If the feedback of the users is large enough a generic model refinement method is also proposed, for improving the entire performance of the system in the retrieval task. In this case, we refine the deep CNN weights based on the feedback gathered from multiple users, forming a new training set. Experimental results denote the effectiveness of the proposed method in accomplishing better retrieval results with respect to a certain user's information need. The validation of the proposed approach on the queries used in the relevance feedback as well as on a generic unseen query dataset, demonstrates the improvement on the retrieval performance, as shown in the experimental results.

CCS Concepts

•Information systems → Users and interactive retrieval; Image search; •Computing methodologies → Neural networks;

Keywords

Relevance Feedback, Content based Image Retrieval, Deep Convolutional Neural Networks

1. INTRODUCTION

Relevance Feedback (RF) is a powerful technique, initially developed during 1960s to improve text-based information retrieval systems [14]. In general, RF refers to the ability of users to impart their judgement regarding the relevance of search results to the system. Then, the system can use this information to ameliorate its performance through iterative steps. The concept of RF was introduced in content based image retrieval (CBIR) in [15], as a technique which can be used in order to bridge the gap between high level semantic concepts in user's mind and low level image features.

Numerous RF approaches have been proposed in the literature [21]. The vast majority of RF techniques are based on query refinement in order to better represent the user's information need. For instance, the Query Vector Modification method reformulates query vector as the mean difference vector between relevant images and irrrelevant ones, in an attempt to move it towards good examples and away from bad examples [9][16]. Another baseline approach in RF is the reweighting method that changes the distance metric in order to move relevant images, as denoted by the user, closer to the query [3].

From a different viewpoint, Bayesian inference methods use a Bayesian framework in order to estimate the posterior probability that an image from the search dataset is relevant to the query image, given the feedback history [5].

Among the proposed RF methodologies, the most relevant to this work are those using pattern classification techniques that try to separate the classes of relevant and irrelevant images. Some of the most successful learning based techniques are using Support Vector Machines [19], Decision Trees [12], Random Forests [2] or boosting techniques [18].

In this paper a novel RF approach is proposed that uses the state-of-the-art deep CNNs for image retrieval. The proposed idea is to use the ability of a deep CNN to modify its internal structure in order to produce better image representations used for the retrieval based on the feedback of the user. To this end, we adapt the deepest neural layers of the CNN model employed for the feature extraction, so that the feature representations of the images that qualified as relevant by the user come closer to the query representation, while the irrelevant ones move away from the query. We note that the proposed method is generic and can be adapted to be used additionally to the traditional RF approaches, in which we aim to modify the query representation in order to come closer to the relevant representations and to move away from the irrelevant ones. We have not proceeded in this paper in the combination of these approaches since we wanted to highlight the power of the deep CNN in the RF framework and the extensions are in most cases straightforward.

The proposed approach overcomes the main drawback of the learning based RF methods, which is the small accuracy caused by the small training set, exploiting the CNN's transfer learning ability. That is the ability of a system to apply

SETN '16, May 18-20, 2016, Thessaloniki, Greece

DOI: http://dx.doi.org/10.1145/2903220.2903240

knowledge learned in previous tasks, since the deep CNN is already trained in millions of images in order to learn generic image representation [6], to a novel task which is defined by the specific query or queries.

We also propose a method for generic deep CNN model refinement, based on multiple users' feedback, in order to improve the CBIR system in a more permanent manner. For this, we gather information from multiple users and then we fine-tune the weights of the pre-trained neural model so that relevant images for each user query come closer to the certain query, while the irrelevant ones move further away. In this case, multiple targets may be defined for each image that are handled automatically by the error back-propagation learning used. The entire framework supports fast learning and inference using parallel processing in the Graphics Processing Unit (GPU) using the CaffeNet framework, described in Section 3, rendering the proposed approach amenable for large-scale and possibly big data retrieval. However, we have not currently proceeded in evaluating the proposed approach in big data which is the direction in future research.

The remainder of the manuscript is structured as follows: Deep CNNs and their applications in CBIR are presented in Section 2. The proposed RF approach in deep CNNs is described in detail in Section 3. The proposed approach for the CNN model refinement in CBIR is presented in Section 4. Experimental results are provided in Section 5. Finally, conclusions are drawn in Section 6.

2. DEEP CNNS AND THEIR APPLICATIONS IN CBIR

In the last few years, deep CNNs have been proven to be very efficient in many vision recognition tasks, such as image classification [10], digit recognition [4][11], and pedestrian detection [17]. As a consequence, they attracted the interest of the computer vision research community as well as the interest of the biggest companies that deal with semantic visual content analysis, description and retrieval. Deep CNNs belong to Deep Learning algorithms which attempt to model high-level abstractions in data by employing deep neural network architectures composed of multiple non-linear transformation neural layers, imitating the human brain that is organized in a deep architecture and processes information through multiple stages of transformation and representation [7].

Deep CNNs, comprise of a number of convolutional and subsampling layers with nonlinear neural activations, followed by fully connected layers. That is, the input image is introduced to the neural network as a 3 dimensional tensor with dimensions equal to the dimensions of the image and the number of color channels (usually 3 in RGB images). Three dimensional filters are learned and apllied in each layer where convolution is performed and the output is passed to the neurons of the next layer for nonlinear transformation using appropriate activation functions. After multiple convolution layers and subsampling the structure of the deep architecture changes to fully connected layers and single dimensional signals. These activations are ususally used as deep representations for classification, clustering or retrieval. Training multiple layers of convolutional kernels can lead to complex features for object recognition achieving superior performance over hand-crafted image descriptors like SIFT, HOG, VLAD, etc. Additionally, it has been shown, that features extracted from the activation of a CNN trained in a fully supervised fashion on a large, fixed set of object recognition tasks can be re-purposed to novel generic recognition tasks [8].

Inspired by these results, deep CNNs introduced in the related problem of CBIR [1][20]. The main approach of applying deep CNNs in the retrieval domain is to obtain feature representations from a pre-trained model, by simply passing images through the model and taking activation values usually drawn from the last layers, which are meant to contain high-level semantic information. As a way of improving the retrieval performance, it has also been proposed the retraining of the convolutional architecture on a dataset with relevant image statistics and classes to the dataset considered at test time [1]. Another proposed technique in the direction of enhancing retrieval perfomance is the refinement of the pre-trained model parameters with class information [20]. Finally, a different proposed approach uses spatial search with CNN features for the retrieval task [13].

3. RELEVANCE FEEDBACK IN DEEP CNN

In this paper we propose a CBIR method with a relevance feedback mechanism, based on deep CNN representations. We utilize the BVLC Reference CaffeNet model[1], which is an implementation of the AlexNet model [10] trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 to classify 1000 ImageNet classes [6]. The model consists of eight trained neural network layers; the first five are convolutional and the remaining three are fully connected. Max-pooling layers follow the first, second and fifth convolutional layers, while the ReLU non-linearity ($f(x) = max(0, x)$) is applied to every convolutional and fully connected layer, except the last fully connected layer (denoted as FC8). The output of FC8 layer is a distribution over 1000 ImageNet classes. The softmax loss is used during the training.

We use the CaffeNet model, either to directly extract representations from the 7th neural network layer (denoted as FC7) or to retrain it on a specific dataset. In the second case, all layers of the new model are initialized with CaffeNet model parameters, except the FC8 which is replaced by a new classification layer that represents the labels of the given dataset and is initialized randomly. All the convolutional layers up to the first fully connected layer (denoted as FC6) remain unchanged and we update FC7 and FC8 layers using error backpropagation. Then, we use the activations of the FC7 layer in order to extract feature representations. The dimension of the FC7 layer is 4096 features.

The above features are calculated both for the set of images to be searched and the query image. Given the query and its corresponding feature representation, the output of the CBIR procedure includes a search in the feature space, in order to find a set of images whose feature representations are closer in terms of their euclidean distance to the query representation. Then, in the first RF round, the user marks some of the retrieved images as relevant or irrelevant and provides his judgement as feedback to the system. Subsequently, this feedback is used by the system to update the weights of the CNN model, so that relevant representations come closer to the query representation while the irrelevant

[1]https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

ones move further away. We note that the representations obtained from a CNN model for a set of input images are adjustable by modifying the weights of the model. For this, we adapt the pre-trained model by removing FC8 layer, and then we fine-tune the parameters of FC7 layer by training on relevant and irrelevant images. The euclidean loss is used during training for the regression task. We should note here that instead of modifying the query, as it is done in many RF approaches, in the proposed method we try to modify the image representation in the seventh neural layer, FC7. This approach can be easily combined with simultaneous modification of the query but it is not examined in this paper.

Let us denote by $\mathbf{q} \in \mathbb{R}^{4096\times 1}$ the feature representation emerged in FC7 layer for the query image, $\mathcal{X}^+ = \{\mathbf{x}_i \in \mathbb{R}^{4096\times 1}, i = 1,\ldots,N\}$ the set of feature representations of N images that have been qualified as relevant by the user, and $\mathcal{X}^- = \{\mathbf{x}_j \in \mathbb{R}^{4096\times 1}, j = 1,\ldots,M\}$ the set of M irrelevant feature representations. Then, our goal is to modify the above relevant and irrelevant representations exploiting the retraining abilities of the neural network. The new target representations for the relevant and irrelevant image representations can be respectively determined by solving the following optimization problems:

$$\min_{\mathbf{x}_i\in\mathcal{X}^+} \mathcal{J}^+ = \min_{\mathbf{x}_i\in\mathcal{X}^+} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{q}\|_2^2, \tag{1}$$

and

$$\max_{\mathbf{x}_j\in\mathcal{X}^-} \mathcal{J}^- = \max_{\mathbf{x}_j\in\mathcal{X}^-} \sum_{j=1}^{M} \|\mathbf{x}_j - \mathbf{q}\|_2^2. \tag{2}$$

We solve the above optimization problems using gradient descent. The first-order gradients of the objective functions $\mathcal{J}^+$ and $\mathcal{J}^-$ are given respectively by:

$$\begin{aligned}\frac{\partial \mathcal{J}^+}{\partial \mathbf{x}_i} &= \frac{\partial}{\partial \mathbf{x}_i}\left(\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{q}\|_2^2\right)\\ &= \frac{\partial}{\partial \mathbf{x}_i}\left((\mathbf{x}_i - \mathbf{q})^\intercal(\mathbf{x}_i - \mathbf{q})\right)\\ &= 2(\mathbf{x}_i - \mathbf{q}),\end{aligned} \tag{3}$$

and

$$\begin{aligned}\frac{\partial \mathcal{J}^-}{\partial \mathbf{x}_j} &= \frac{\partial}{\partial \mathbf{x}_j}\left(\sum_{j=1}^{M} \|\mathbf{x}_j - \mathbf{q}\|_2^2\right)\\ &= \frac{\partial}{\partial \mathbf{x}_j}\left((\mathbf{x}_j - \mathbf{q})^\intercal(\mathbf{x}_j - \mathbf{q})\right)\\ &= 2(\mathbf{x}_j - \mathbf{q}).\end{aligned} \tag{4}$$

Consequently, the update rules for the n-th iteration can be formulated as:

$$\mathbf{x}_i^{(n+1)} = \mathbf{x}_i^{(n)} - 2\alpha(\mathbf{x}_i^{(n)} - \mathbf{q}), \quad \mathbf{x}_i \in \mathcal{X}^+ \tag{5}$$

and

$$\mathbf{x}_j^{(n+1)} = \mathbf{x}_j^{(n)} + 2\alpha(\mathbf{x}_j^{(n)} - \mathbf{q}), \quad \mathbf{x}_j \in \mathcal{X}^- \tag{6}$$

where the parameter $\alpha \in [0, 0.5]$ controls the desired distance from the query representation.

Using the above representations at the $n+1$ round as targets in the FC7 layer we formulate a regression task for the neural network and the deep CNN, which is trained using backpropagation, produces representations for the relevant and the irrelevant images as close as possible to the targets $\mathbf{x}^{(n+1)}$. Thereafter the network's convergence the relevant images are closer to the query in the FC7 representation layer and the irrelevant are further away from the query in the FC7 representation layer. Thus, the RF procedure is integrated by feeding the images of the given dataset and the query image into the input layer of the modified model and obtaining the new FC7 representations.

The above process is performed in each RF round, by initializing the CNN model with the parameters of the previous round and re-training on the new set of relevant and irrelevant images with their corresponding updated targets. The above procedure is denoted as RF(FC7). Given a new query from the user, the system executes the procedure from the beginning.

4. DEEP CNN MODEL REFINEMENT IN CBIR

Exploiting the idea that a deep neural architecture can non-linearly distort the feature space in order to bring closer the image representations to a specific query we can consider using user feedback to multiple queries for more permanent CNN modifications. That is, the second proposed method refers to the refinement of the deep CNN model in order to ameliorate the entire system's performance in the CBIR task in a permanent manner, as opposed to the aforementioned proposed method which aims to improve the retrieval performance for a certain user's information need.

The method consists of two phases. In the first phase, the system merely gathers information from different users' feedback and stores it. This information consists of queries and relevant and irrelevant images to these queries. In the second phase, the system builds targets for each image based on the user queries using (5) and (6) and refines either the CaffeNet model or the refined with class labels model, by removing the FC8 layer and training on relevant and irrelevant images gathered in the first phase. The refinement procedure is the same as in the first proposed method, described in the previous section.

An issue arising from using multiple users' feedback is that in the phase of re-training the CNN model with inputs the annotated as relevant or irrelevant images and targets the ones deriving from the equations 5 and 6, an image can be marked in different ways by different users and, thus, be involved in different queries. That is, a given image representation $\mathbf{x}_t$ in FC7 can be relevant to the query $\mathbf{q}_1$ and irrelevant to $\mathbf{q}_2$. This forms two contradicting update rules for $\mathbf{x}_t$. That is: $\mathbf{x}_t^{(n+1)} = \mathbf{x}_t^{(n)} - 2\alpha(\mathbf{x}_t^{(n)} - \mathbf{q}_1)$ and $\mathbf{x}_t^{(n+1)} = \mathbf{x}_t^{(n)} + 2\alpha(\mathbf{x}_t^{(n)} - \mathbf{q}_2)$. Employing the CNNs to this task, saves us from searching an unified target for the duplicates. That is, we define the targets for all the annotated images, we append them in the training set and let the network to converge to the most prominent target during network training.

5. EXPERIMENTAL RESULTS

In this section we present experiments conducted in order to evaluate the proposed methods. We have two im-

age retrieval datasets: the 102 Category Flower Dataset[2], consisting of 8189 images divided into 102 categories and the Inria Holidays Dataset [3], consisting of 1491 images divided into 500 classes. In the following results, we denote as FC7 the feature representations obtained from the FC7 layer of the pre-trained CaffeNet model that serves as the baseline comparison method. We denote as RCL(FC7) the representations obtained from the FC7 layer if we retrain the model with the information that is available from the training set which is the class information for each training image. This is obtained, as explained before, by adding a new layer FC8 with softmax loss and random weights and targets the classes of the dataset. The FC8 layer is removed after the training and the refined FC7 representations are used denoted as RCL(FC7). We denote as RF(FC7) and RCL(FC7)→ RF(FC7) the feature representations emerged in the 7th layer after RF in FC7 and RF in RCL(FC7) as described in the previous Sections.

5.1 Relevance Feedback in Deep CNN

In order to assess the performance of the proposed RF method, we perform experiments on the 102 Category Flower Dataset, considering 6147 images for the search dataset. We randomly select 10 images from the rest of the dataset to perform queries. Each class in our dataset consists of between 20 and 238 images. As relevant is considered an image that belongs to the same class as the query and as irrelevant an image belonging to a different class. We execute 4 RF rounds for each query. At each RF round, we use 19 relevant and equal number of irrelevant images for the model refinement. The parameter a is set to 0.25, and the model is trained for 2 epochs at each RF round.

In Inria Holidays dataset, we consider as search set 991 images and we perfom 15 queries from the residue. Each class in the search set consists of between 1 and 12 images. We execute 3 RF rounds for each query. At each RF round we use 5 relevant and 5 irrelvant images for the model refinement. The model is trained for 2 epochs at each RF round.

We measure the performance of our method at each feedback round in terms of Precision, which is the ratio of relevant images within the top T retrieved images, as shown below, and we compute the average precision obtained over all the performed queries. Average precision is measured in scope $T = 20$ and $T = 50$ for the 102-Flower dataset and in scope $T = 5$ and $T = 10$ for the Inria Holidays dataset.

$$Precision = \frac{n.\ of\ Relevant\ Retrieved\ Images}{n.\ of\ Retrieved\ Images} \quad (7)$$

As we can see in Figures 1-9, the proposed RF approach improves notably the performance of the system both for the FC7 and RCL(FC7) representations by the first RF round.

5.2 Deep CNN Model Refinement in CBIR

In the second set of experiments conducted in order to evaluate the method of the CNN model refinement in CBIR, we employ the 102 Flower dataset, with search set of 6147 images and a set of 102 different queries (queries-set). In order to evaluate the model's generalization ability, we also test the performance on a dataset consisting of 1938 additional unknown queries (generic-set). For each different user, we employ 19 relevant and 19 irrelevant images that form a train set consisting of 3876 images.

In the Inria Holidays dataset we consider a search set of 991 images and a set of 500 different queries. For each different user we use 1 relevant and 5 irrelevant images that form a train set of 3000 images.

For the evaluation we use the mean Average Precision, which is the the mean value of the Average Precision (AP) of all the queries. The AP is defined as the average of precision values computed at the point of each correctly retrieved

[2]http://www.robots.ox.ac.uk/vgg/data/flowers/102/
[3]https://lear.inrialpes.fr/~jegou/data.php

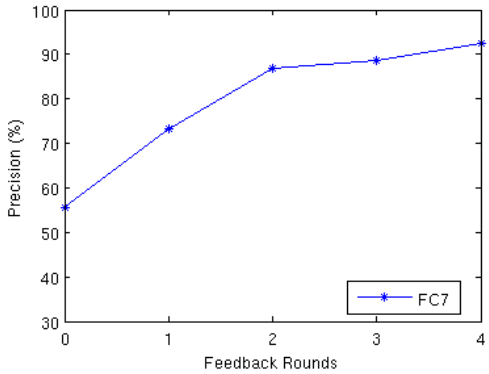

Figure 1: 102-Flower - FC7, T=20

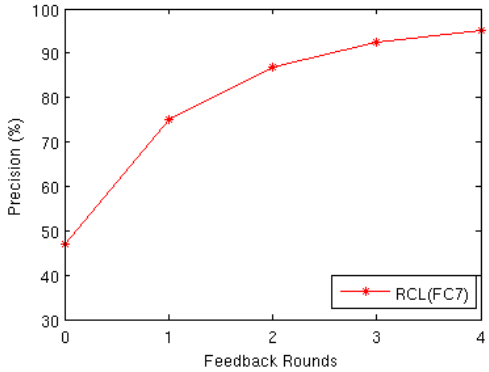

Figure 2: 102-Flower - RCL(FC7), T=20

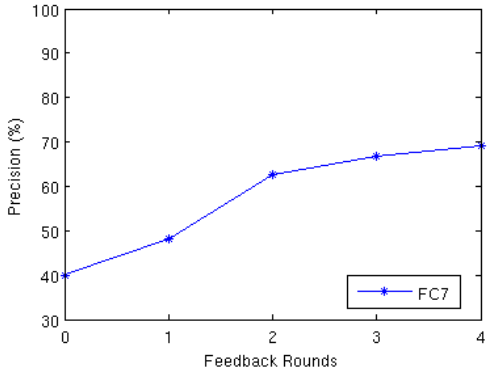

Figure 3: 102-Flower - FC7, T=50

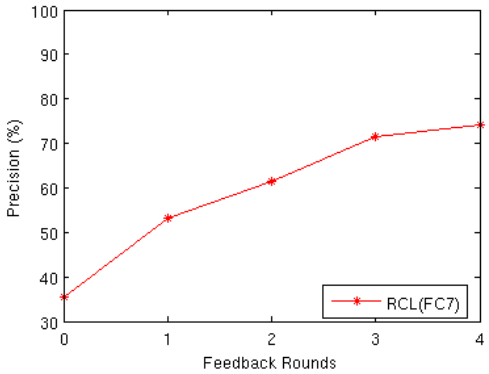

Figure 4: 102-Flower - RCL(FC7), T=50

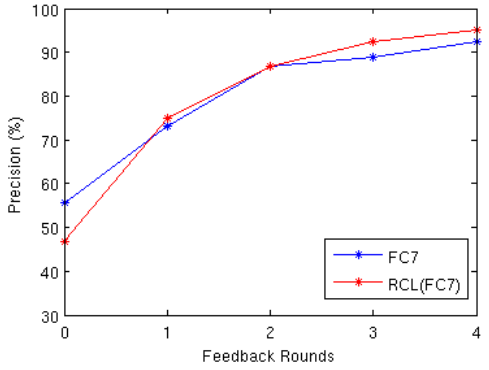

Figure 5: 102-Flower - FC7 - RCL(FC7), T=20

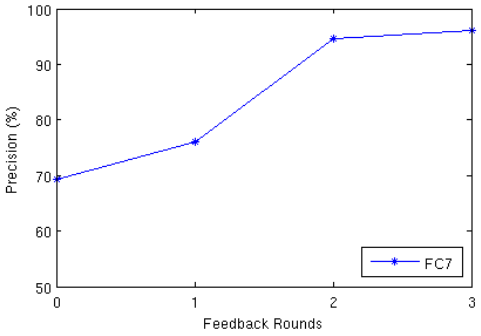

Figure 6: Inria - FC7, T=5

item. This is formulated for the i-th query as follows:

$$AP_i = \frac{1}{Q_i} \sum_{n=1}^{N} \frac{R_i^n}{n} t_n^i, \tag{8}$$

where Q_i is the total number of relevant images for the i-th query, N is the total number of images of the search set, R_i^n is the number of relevant retrieved images within the n top results; t_n^i is an indicator function with $t_n^i = 1$ if the n-th retrieved image is relevant to the i-the query, and $t_n^i = 0$ otherwise.

Results are demonstrated in Tables 1 and 2. We denote as $mAP1$, the mean Average Precision values of the queries-set,

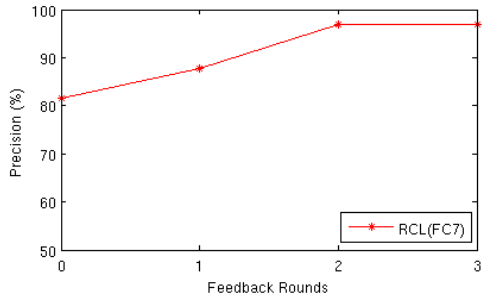

Figure 7: Inria - RCL(FC7), T=5

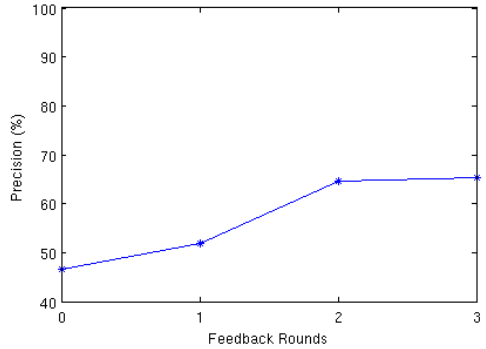

Figure 8: Inria - FC7, T=10

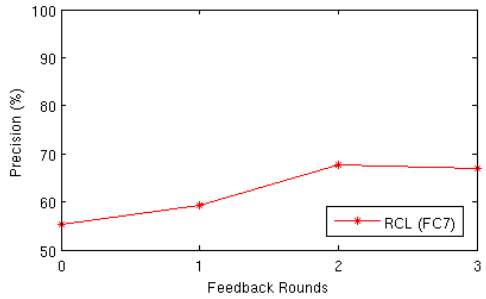

Figure 9: Inria - RCL(FC7), T=10

and as $mAP2$ the mean Average Precision of the generic-set. We also demonstrate the Precision-Recall curve for the top $T = 200$ and $T = 991$ retrieved images in the 102-Flower and the Inria Holidays dataset respectively. The recall metric is defined as follows:

$$Recall = \frac{n.\ of\ Relevant\ Retrieved\ Images}{n.\ of\ Relevant\ Images} \tag{9}$$

From the illustrated results, we observe that the proposed method improves the system's performance in the retrieval task in any considered case. Furthermore, we can see that the RCL(FC7) representations perform better in the CBIR task with respect to the FC7 representations, and so do the RCL(FC7) $\rightarrow$ RF(FC7) representations as compared to the RF(FC7). Finally, it is interesting to see that in the experiments on the 102 Flower dataset the system's performance has been improved not only on the set of the queries involved in the CNN model refinement, but also on the generic

set. That confirms the generalization ability of the proposed method.

Table 1: Flower-102: mAP

Feature Representation	mAP 1	mAP 2
FC7	0.3140	0.3201
RF(FC7)	0.4770	0.3830
RCL(FC7)	0.3494	0.3808
RCL(FC7) → RF(FC7)	**0.5323**	**0.4662**

Table 2: Inria Holidays: mAP

Feature Representation	mAP
FC7	0.6989
RF(FC7)	0.7717
RCL(FC7)	0.7353
RCL(FC7) → RF(FC7)	**0.7787**

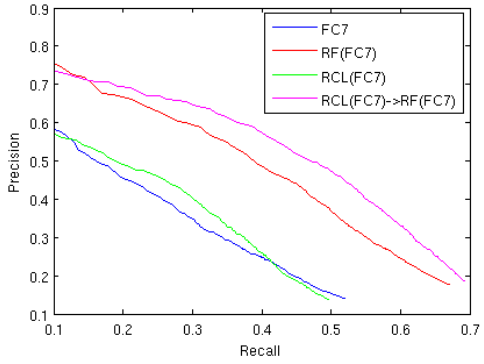

Figure 10: 102 Flower - Precision-Recall curve, Queries-Set

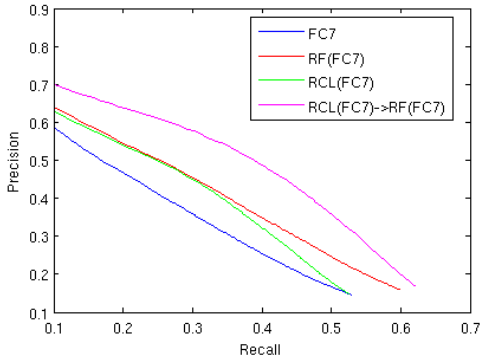

Figure 11: 102 Flower - Precision-Recall curve, Generic-Set

6. CONCLUSIONS

In this paper we proposed two novel techniques that exploit relevance feedback and employ deep Convolutional Neural Network models in the Content Based Image Retrieval task. The first, RF method, uses a deep CNN in order to extract the feature representations of the images and asks for the feedback of the user. Then, the proposed approach improves the retrieval performance for the user's information need, by adapting the CNN model and fine-tuning with relevant and irrelevant images as marked by the user, in order to make relevant representations come closer to the query and the irrelevant ones to move further away. The second, model refinement method, improves the entire performance of the system in the CBIR task by using information gathered from multiple users and refining the weights of the deep CNN model, as to make each relevant representation come closer to its corresponding query and each irrelevant to move away from it. Experimental results on two image datasets illustrate the improvements of the proposed approaches. Future research is directed towards combining query modification and large scale experiments.

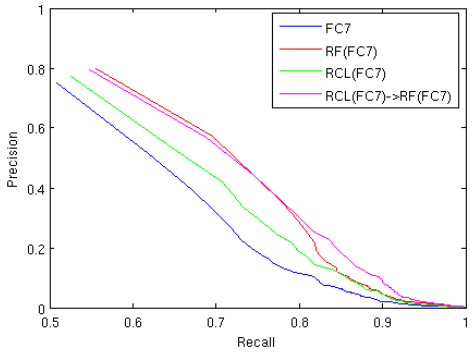

Figure 12: Inria Holidays - Precision-Recall curve

7. REFERENCES

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *Computer Vision–ECCV 2014*, pages 584–599. Springer, 2014.

[2] S. R. Bulo, M. Rabbi, and M. Pelillo. Content-based image retrieval with relevance feedback using random walks. *Pattern Recognition*, 44(9):2109–2122, 2011.

[3] G. Ciocca and R. Schettini. A relevance feedback mechanism for content-based image retrieval. *Information processing & management*, 35(5):605–632, 1999.

[4] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

[5] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Target testing and the pichunter bayesian multimedia retrieval system. In *Digital Libraries, 1996. ADL'96., Proceedings of the Third Forum on Research and Technology Advances in*, pages 66–75. IEEE, 1996.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[7] L. Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, 3:e2, 2014.

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[9] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. *Computer Science Department*, page 551, 1998.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276, 1995.

[12] S. D. MacArthur, C. E. Brodley, and C.-R. Shyu. Relevance feedback decision trees in content-based image retrieval. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 68–72. IEEE, 2000.

[13] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.

[14] J. J. Rocchio. Document retrieval system-optimization and. 1966.

[15] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644–655, 1998.

[16] S. Santini and R. Jain. Integrated browsing and querying for image databases. *MultiMedia, IEEE*, 7(3):26–39, 2000.

[17] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3626–3633. IEEE, 2013.

[18] K. Tieu and P. Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2):17–36, 2004.

[19] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001.

[20] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the ACM International Conference on Multimedia*, pages 157–166. ACM, 2014.

[21] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 8(6):536–544, 2003.